Cracking Open the Mystery of Large-Scale AI

NAIRR Pilot Inaugural Meeting 2025

David Bau
Northeastern University

Gathering to build infrastructure post-covid

A “grand mission”

- A little lab cluster

Belief that it is possible

- Not systems experts!

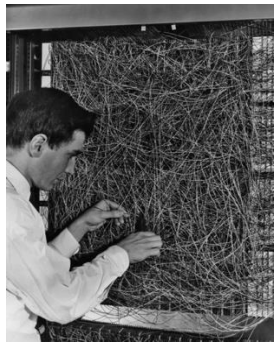
The power to cooperate

- As a team, done in a day.
- Collaborations continue.

an *opportunity* to build an AI research **ecosystem**

Advances in AI have been genuinely surprising

1962, Perceptron: learning



64 params

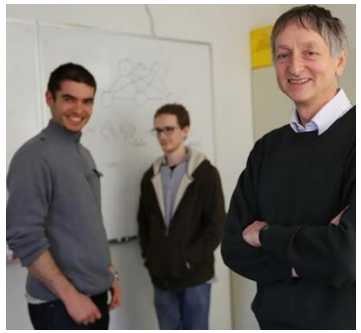
1000x

1986, LeNet: generalization

60,000 params

1000x

2012, Alexnet: SOTA



60,000,000 params

3000x

2020 GPT: metalearning

175,000,000,000 params

3000x

2025 : reasoning

scale up inference iterations

The interpretability question

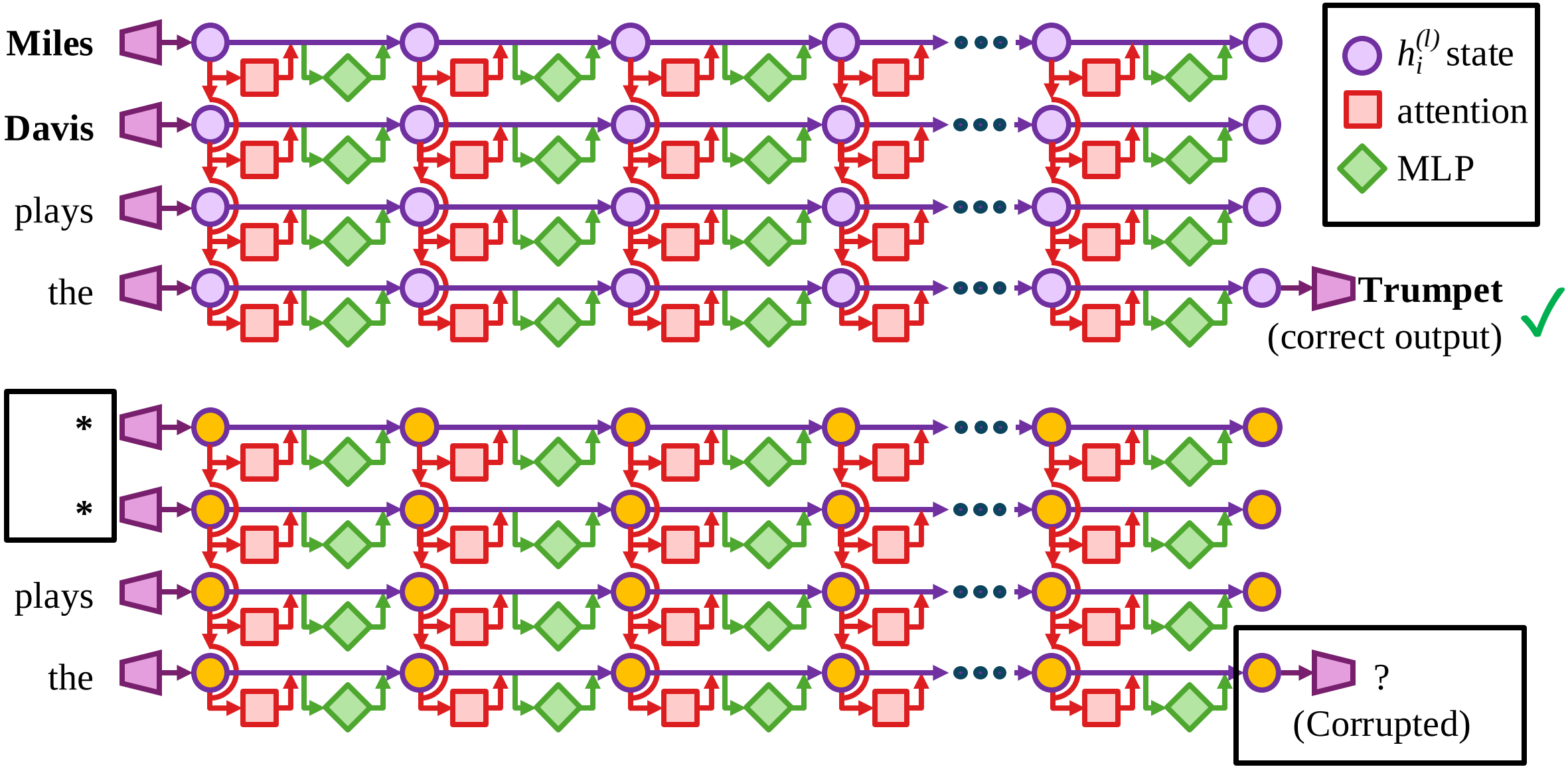
After we train a **large-scale AI**...

...is it possible for humans to **understand** it?

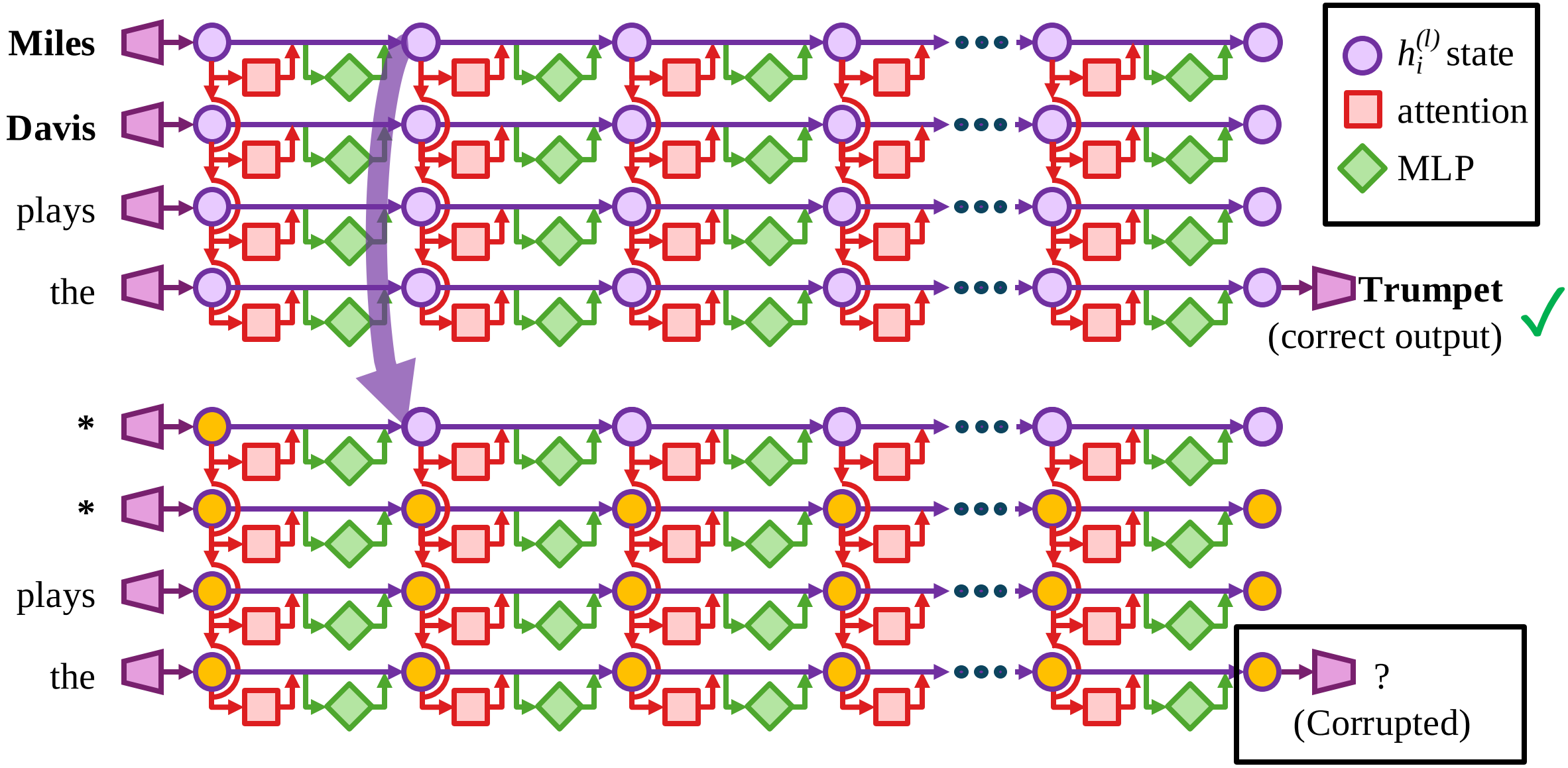
First ingredient for interpretability in AI

1. Understanding AI

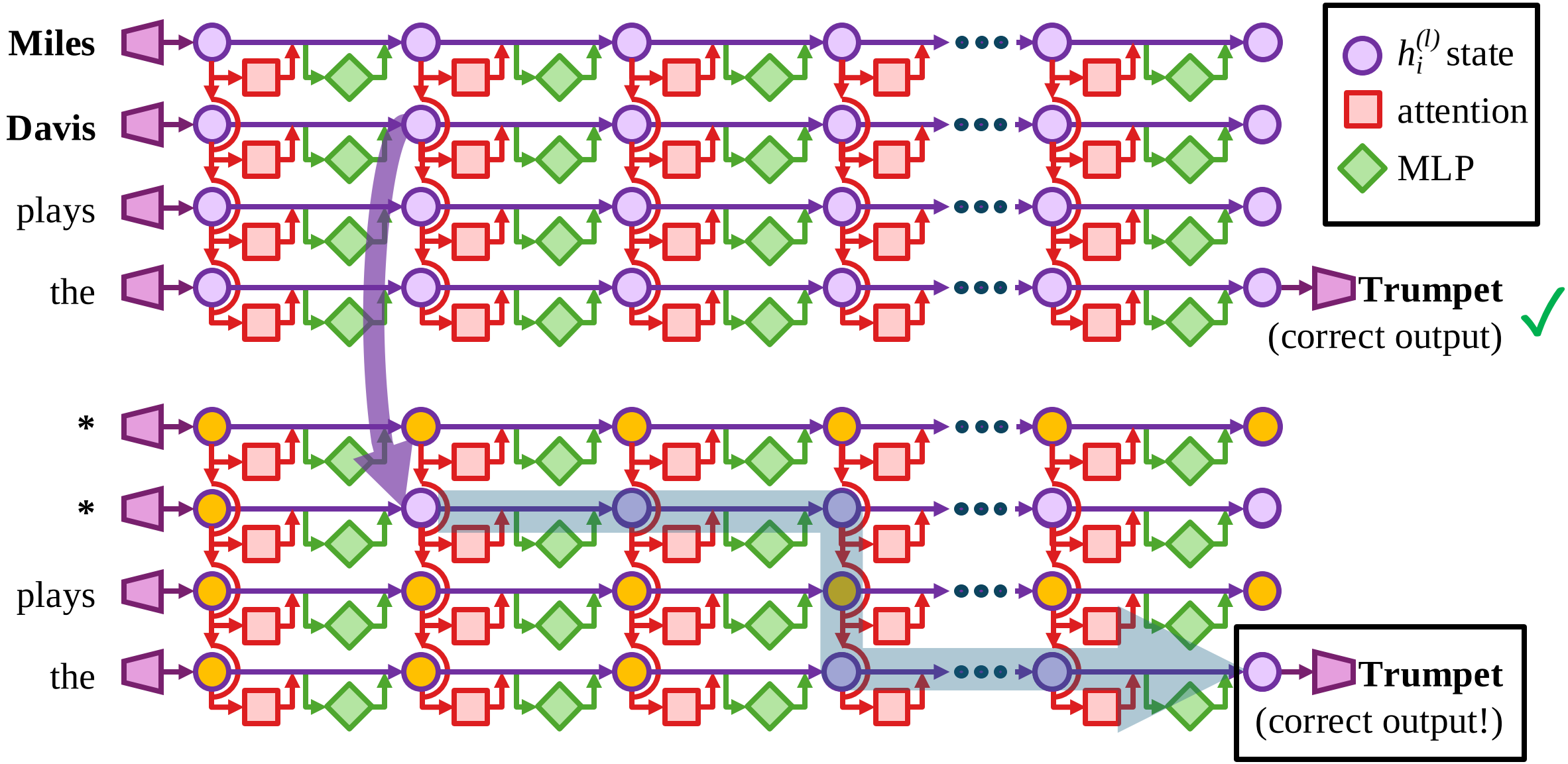
ROME: Understanding LLM Knowledge



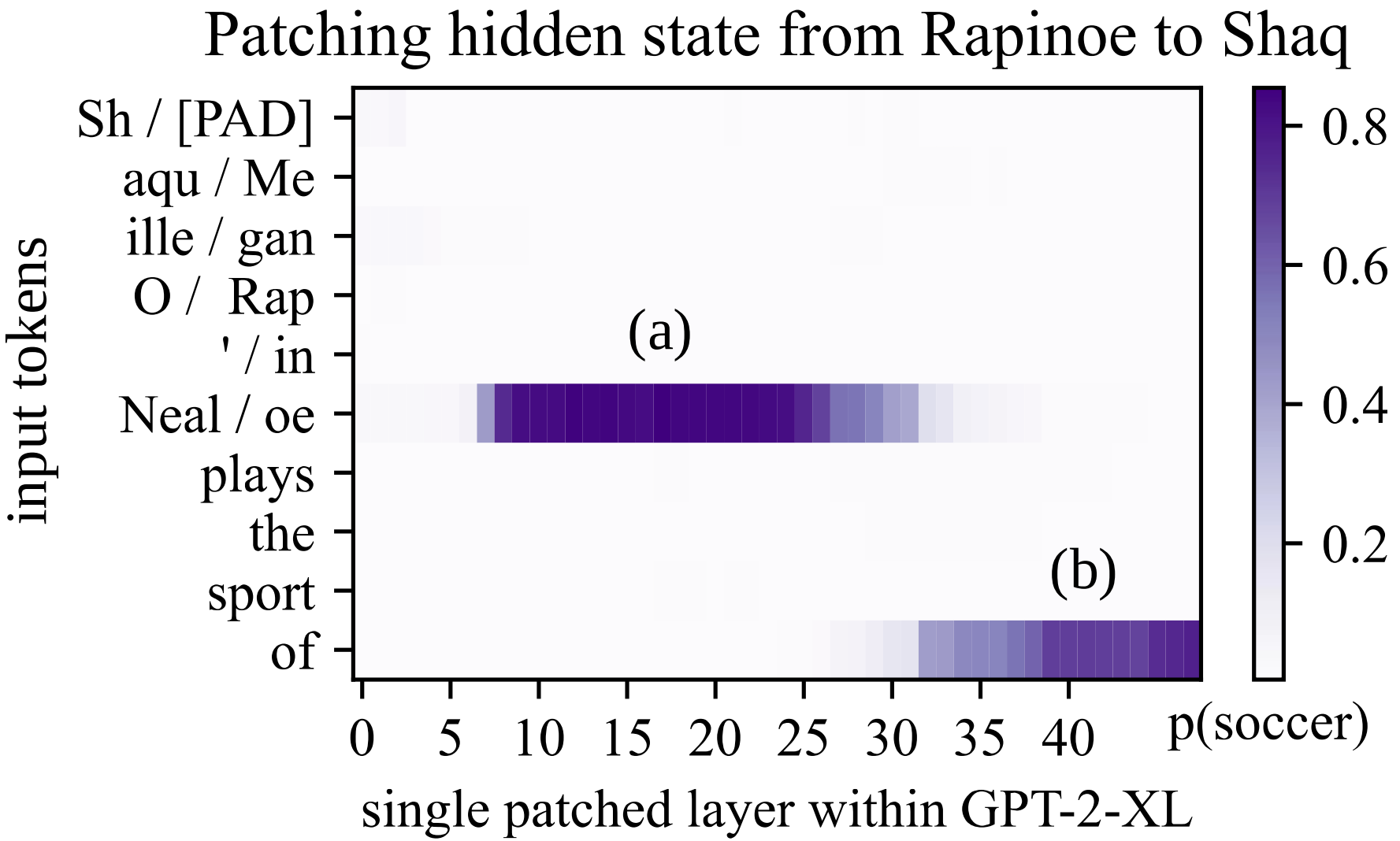
Transplant Hidden State



Transplant Hidden State



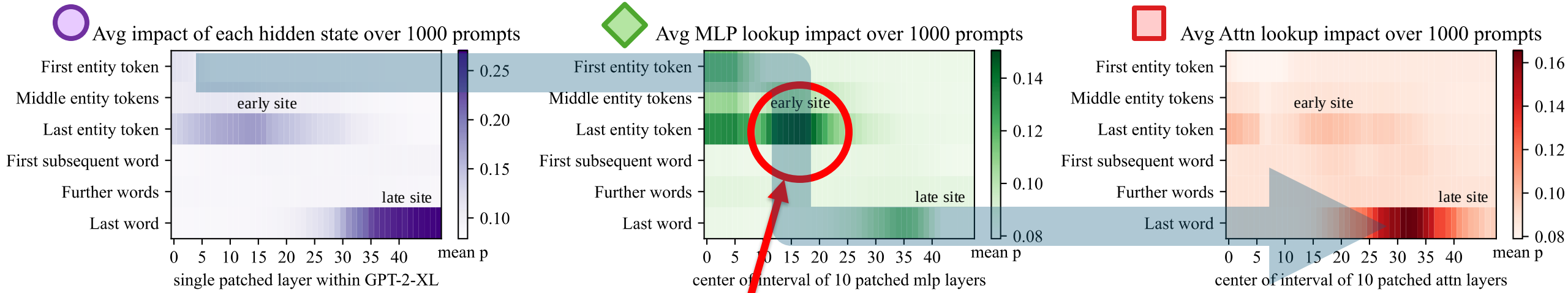
Results: Causal tracing



Copying a state between two sentences:

Now Shaq plays soccer.

ROME: Finding “What Factual Knowledge Is”

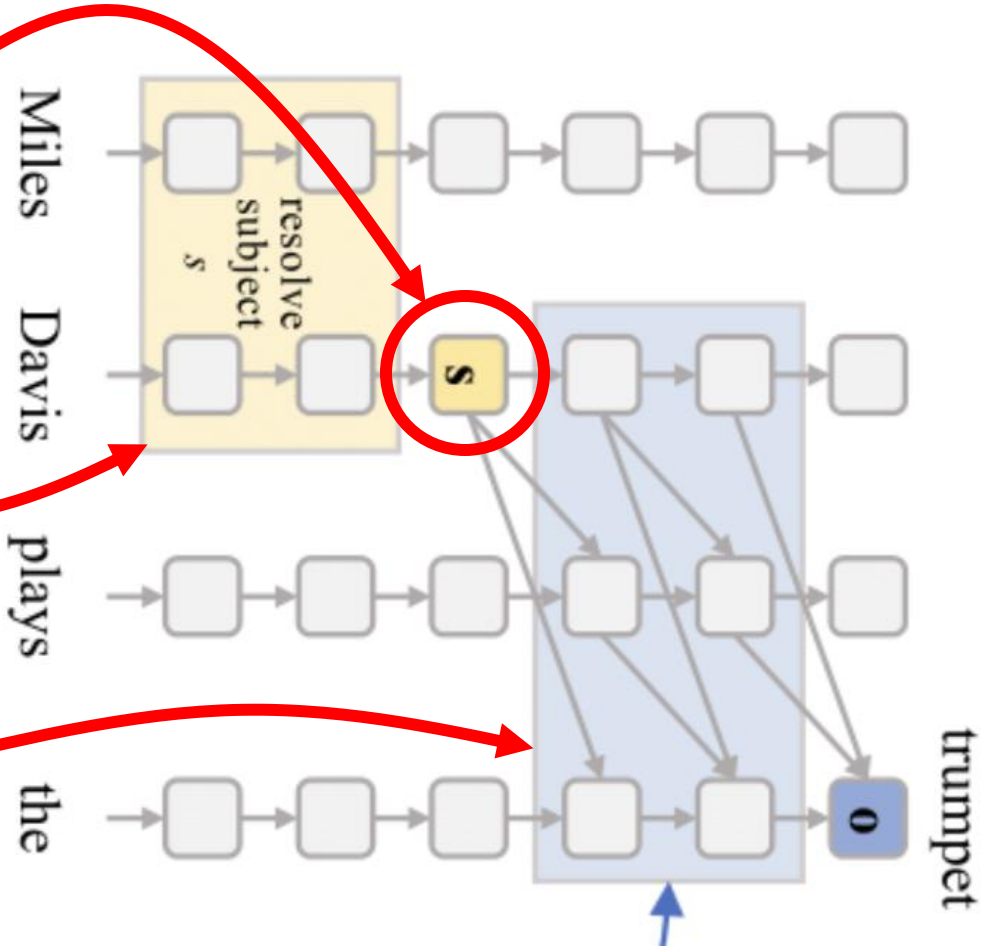


Finding 1: **“Knowledge” is a localized mapping** from early-layer “word/phrase” vectors to late-layer “meaning” vectors

[Meng 2022, “Locating and Editing Factual Associations”]

The ROME view of Understanding AI

1. What “information” is encoded in **individual vectors**?
   “properties of Miles Davis”
2. What ”knowledge” is stored in **vector maps**?
   “Miles Davis plays trumpet”
3. What “reasoning” is embodied in **vector pathways**?
   “say the instrument they play”



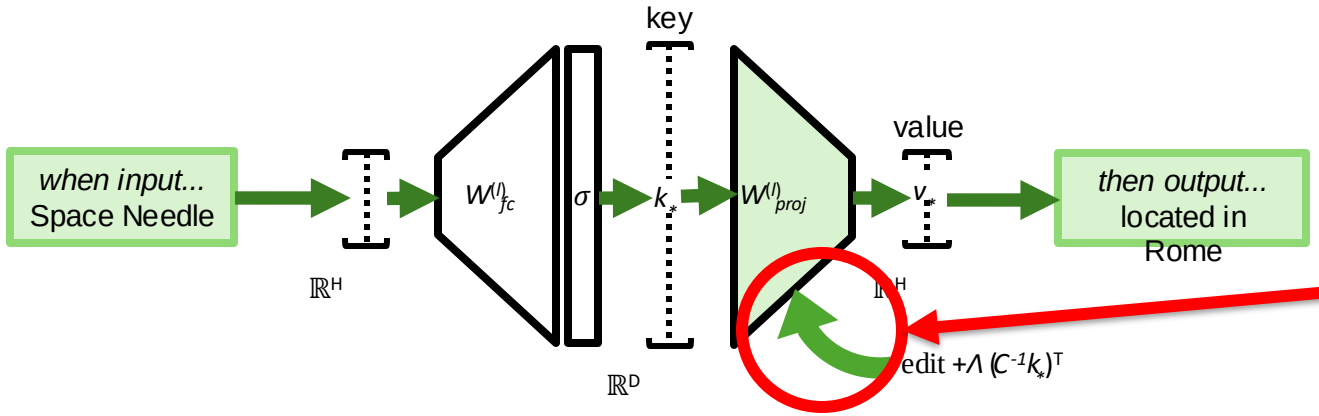
[Geva 2023, “Dissecting Recall of Factual Associations”]

[Hernandez 2024, “Linearity of Relation Decoding”]

Second ingredient for interpretability in AI

2. Controlling AI

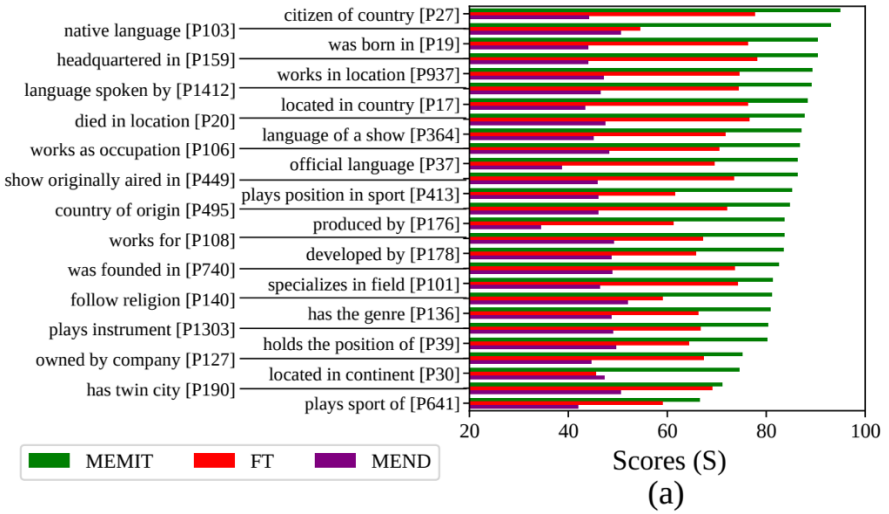
ROME: Editing "How Facts are Mapped"

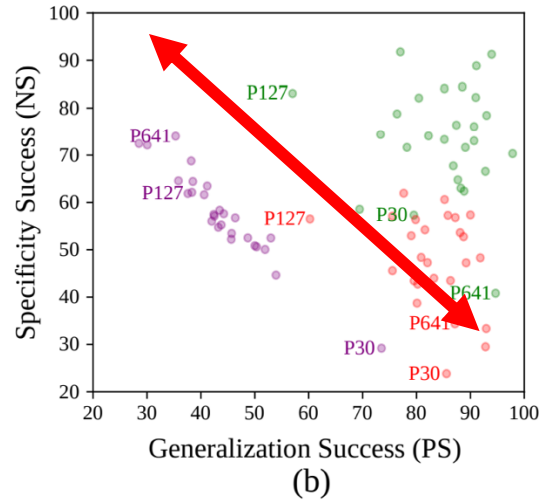


"Knowledge" can be edited by directly editing vector mappings

Finding 2:

With specificity and generalization better than fine-tuning



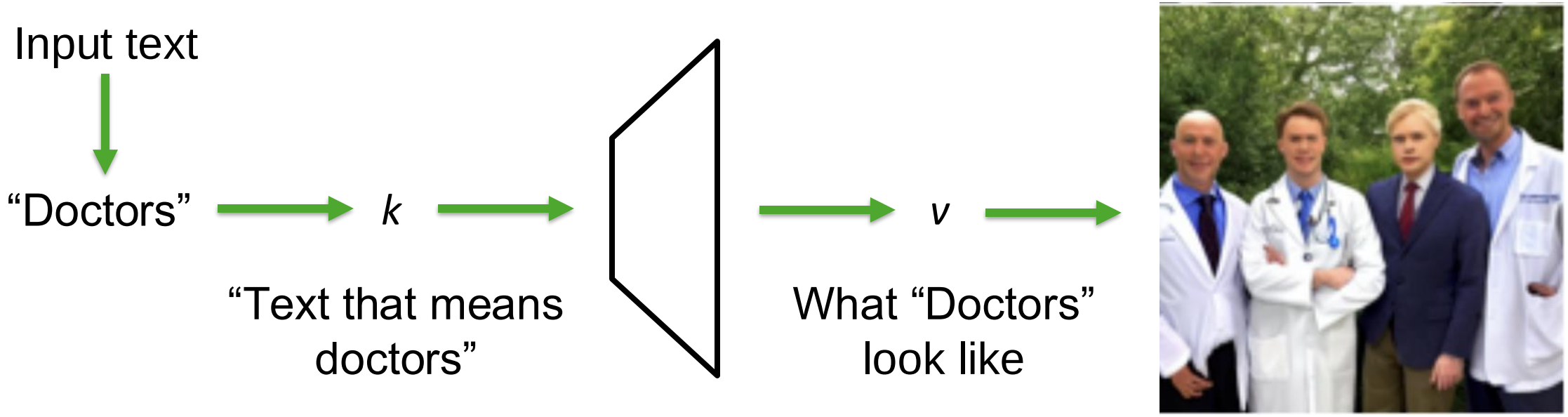


[Meng 2022, "Locating and Editing Factual Associations"]

[Also see: Meng 2022, "MEMIT"]

Applying ROME to an Immediate Problem

Inside a diffusion model, a mapping from text to vision:

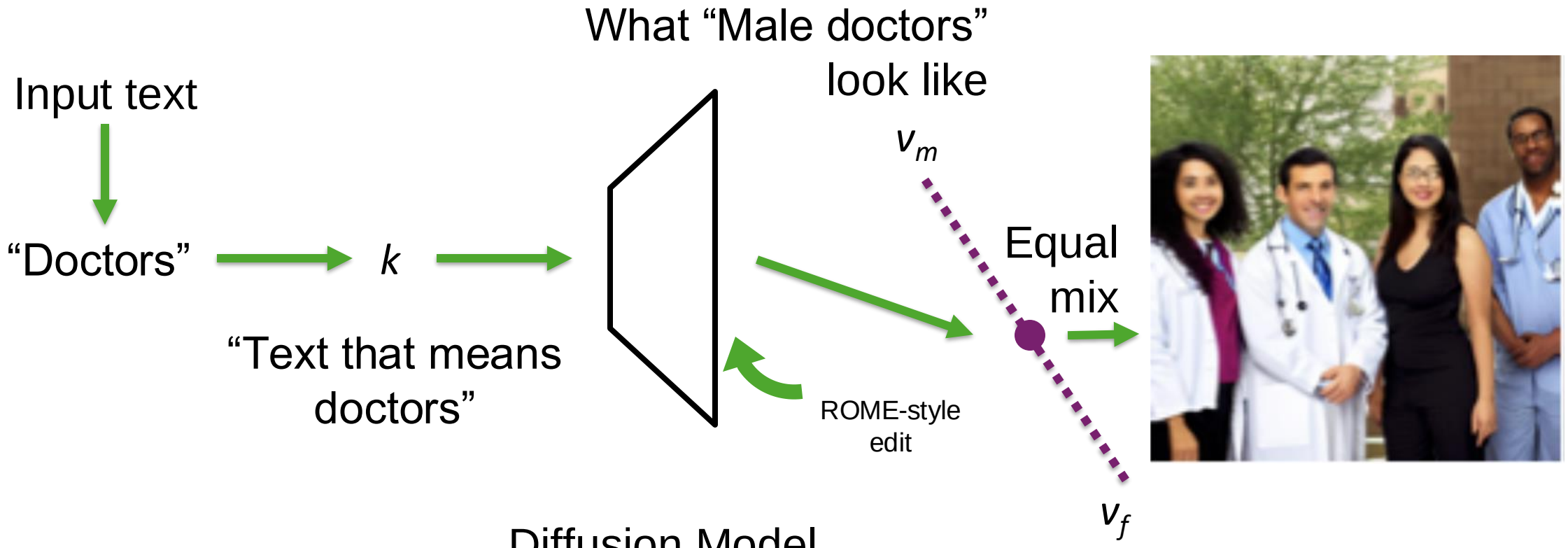


Diffusion Model
Cross-Attention Maps
From Text to Vision

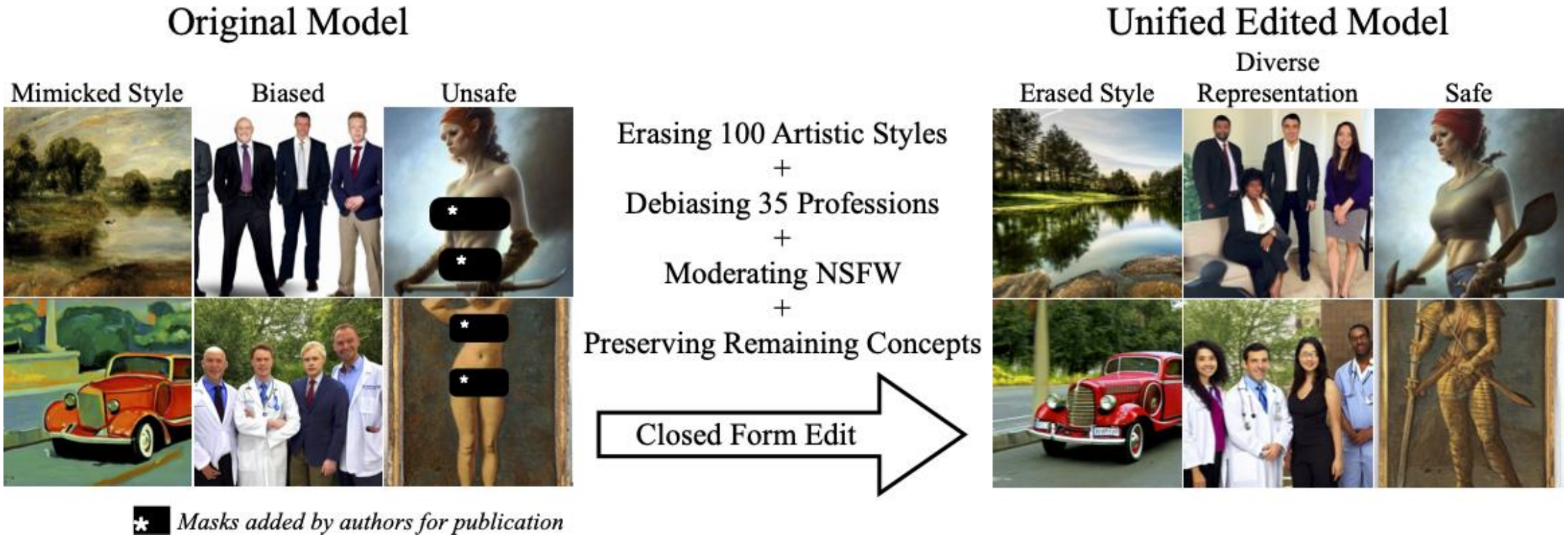
[Gandikota 2024, “Unified Concept Editing”]

Applying ROME to an Immediate Problem

Inside a diffusion model, a mapping from text to vision:



Debugging Many Diffusion Model Problems



Direct Model Editing is fast, so it gives humans a much better "debugging interface" for AI than ordinary fine-tuning.

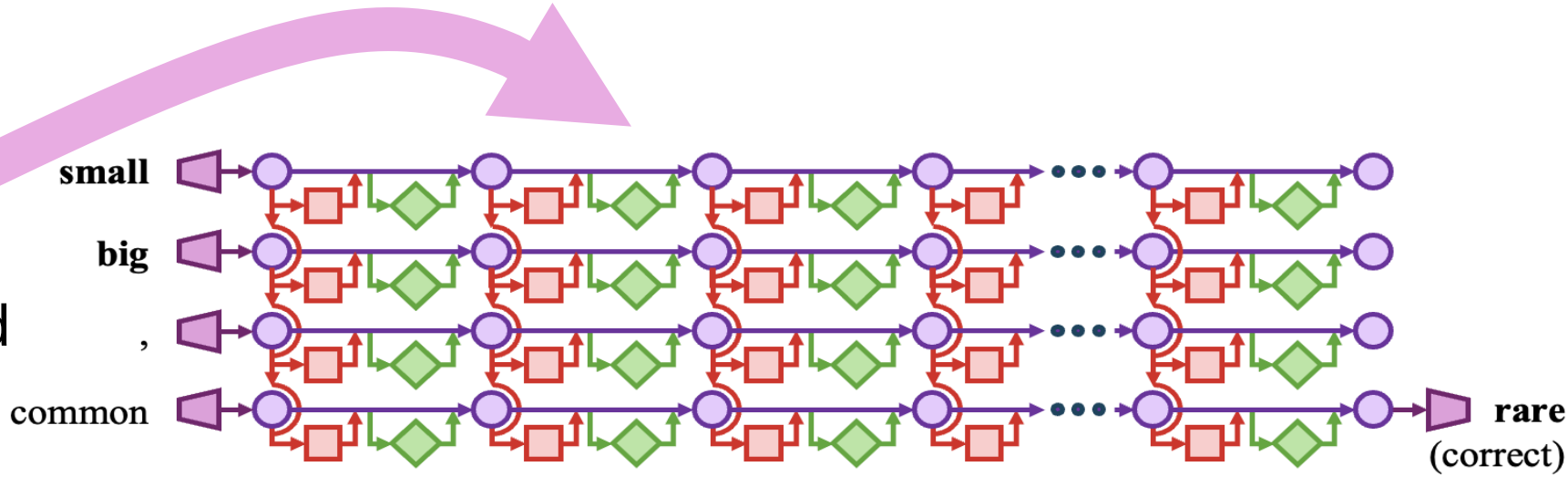
Third ingredient for interpretability in AI

3. Power to muck with AI

The closed tech structure of modern AI

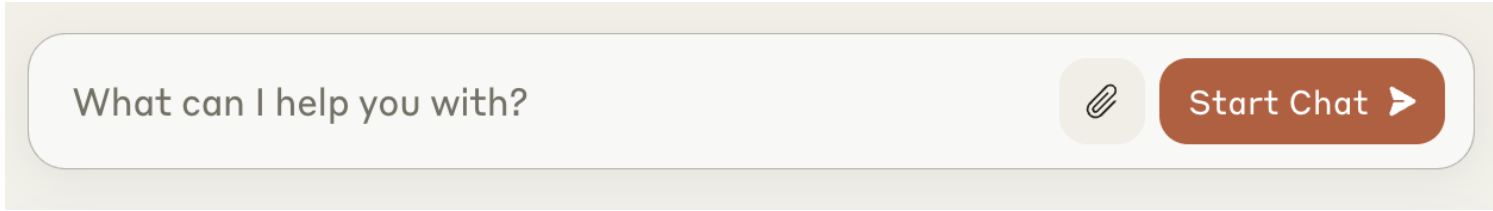
Large-scale AI models are made of two parts:

1. The “foundation”
   hugely expensive engineering that needs to absorb and learn all the world’s information



foundation training electricity: $100,000,000

2. “Interpretation+control”
   small bit of engineering with all the user-facing features and choices: dialog, biases, safety...

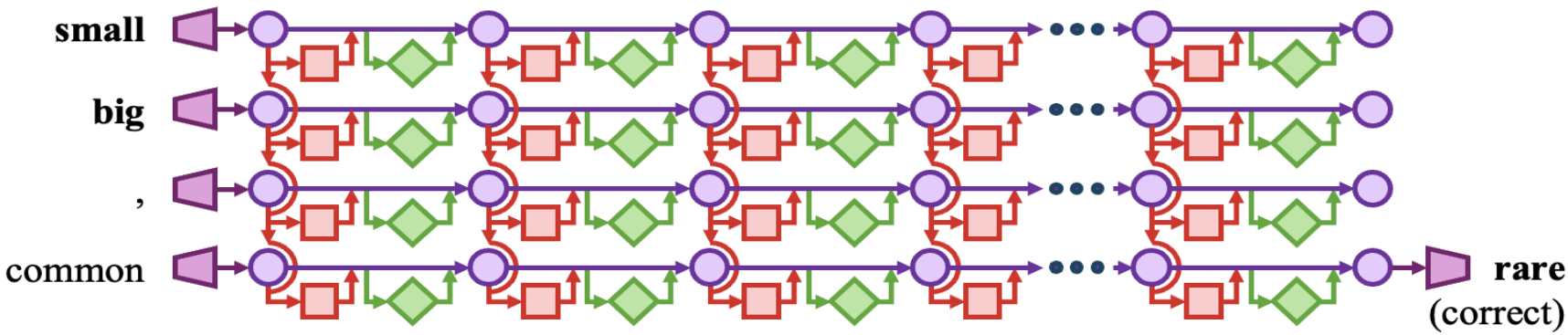


customization electricity: $1,000

Even “Open” models are Closed due to $$

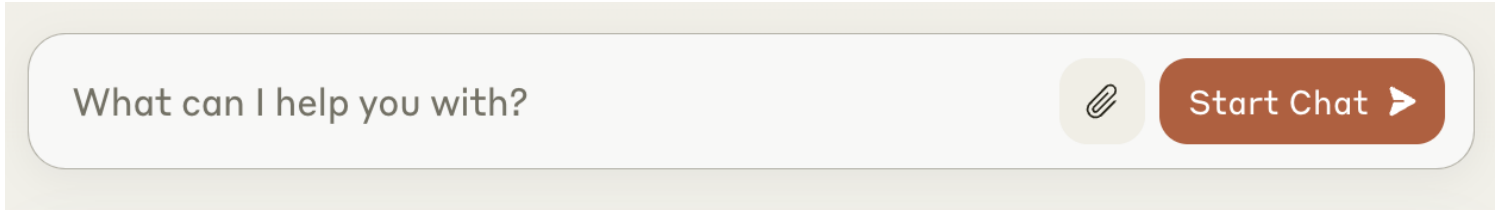
What experiments will you want to do on Llama3-405b?

1. The “foundation”
   needs 1.6 terabytes GPU RAM to run inference



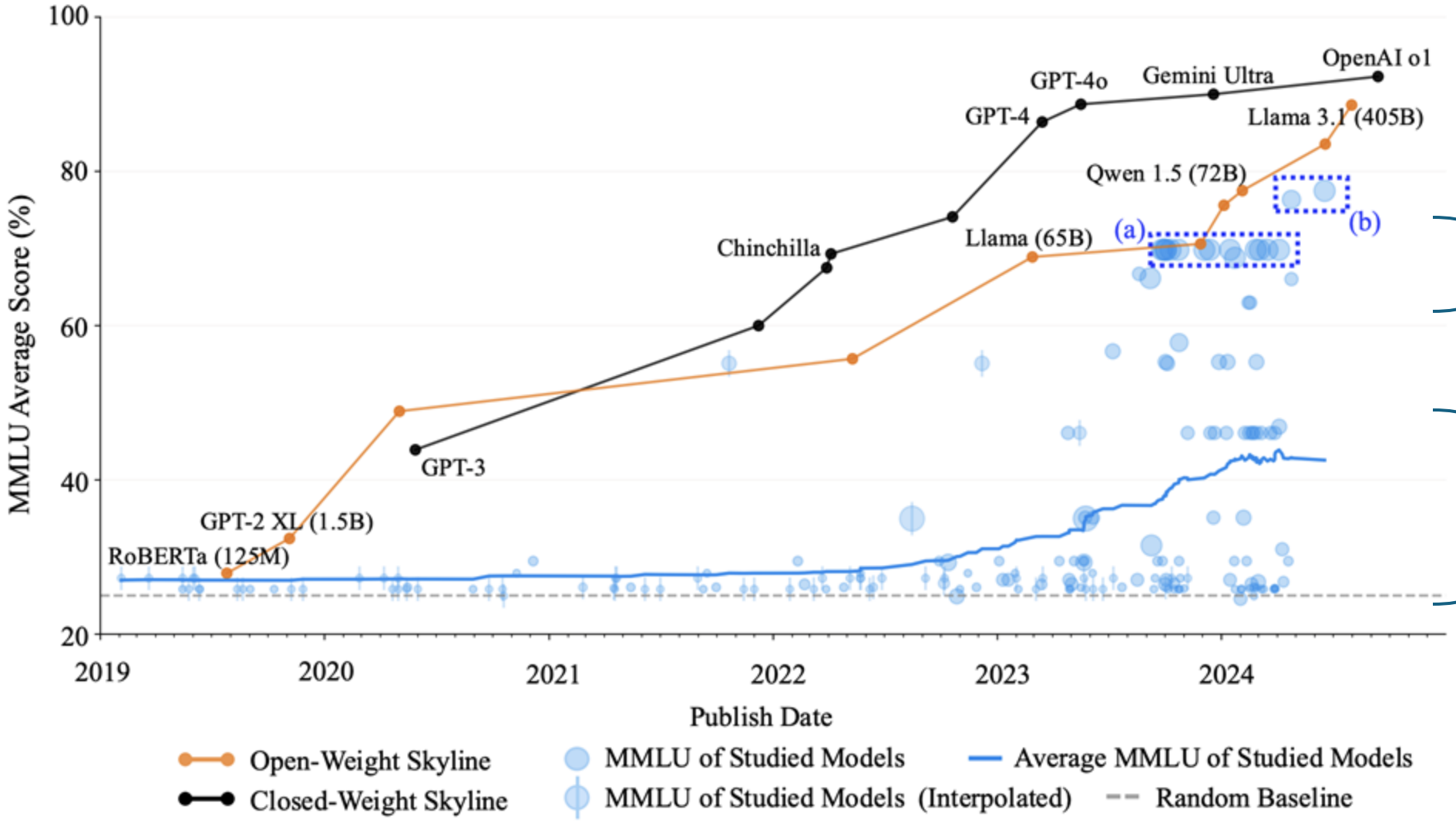
foundation inference capital cost: $2,000,000

2. “Interpretation+control”
   need 1.6 terabytes GPU RAM to run
   ...**there is no discount!**



interpretability lab capital cost: $2,000,000

But we are failing to study big AI



A few studies at the one-GPU limit: 70b params

In 2024 almost all interpretability research is on small models that have very low capabilities

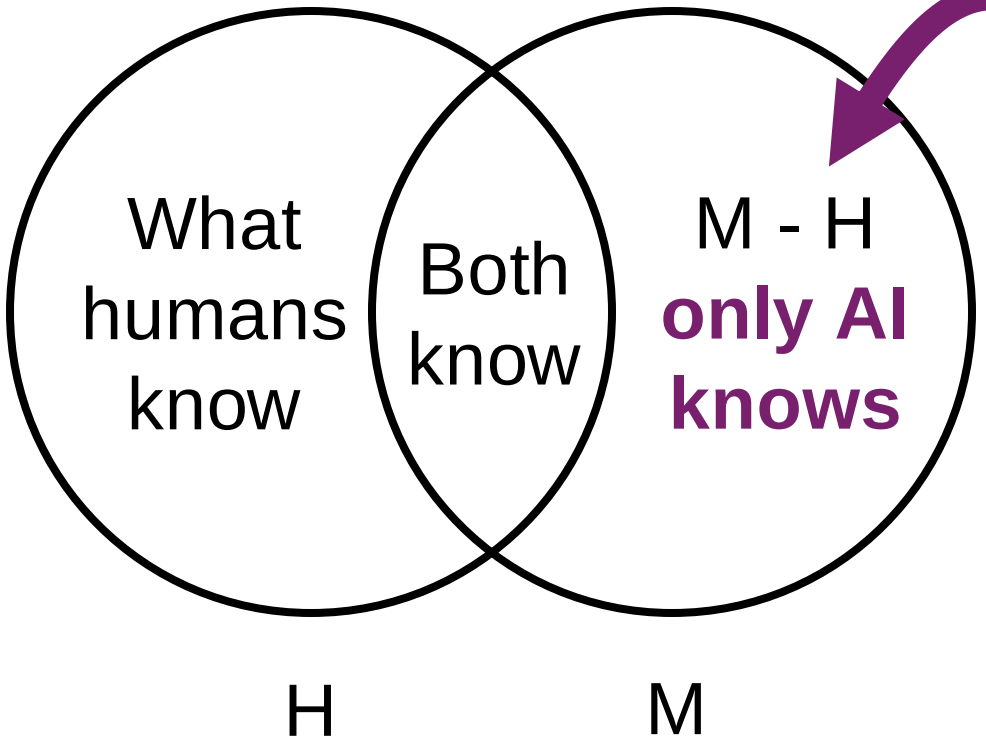
"NNsight and NDIF: Democratizing Access to Foundation Model Internals" Fiotto-Kaufman

Why is this important?

Profound AI knowledge

Unspoken AI **goals**

AIs already know things that humans do not



White to play.

Human chess experts would take this knight and press kingside.

AlphaZero knows that it is better for white to start regrouping queenside.

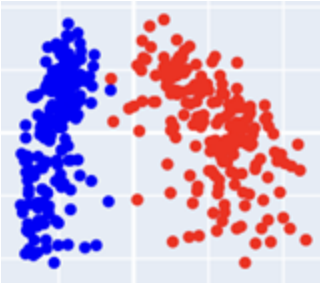
[Lisa Schut, "**Bridging the Human–AI Knowledge Gap**"]

To understand AlphaZero, crack it open

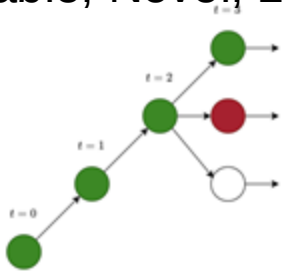
[Lisa Schut, “**Bridging the Human–AI Knowledge Gap**”]

1. Decompose the AI into vector “concepts”
2. Sort the important concepts
3. Teach those concepts to humans

Sparse vectors



Desirable, Novel, Learnable



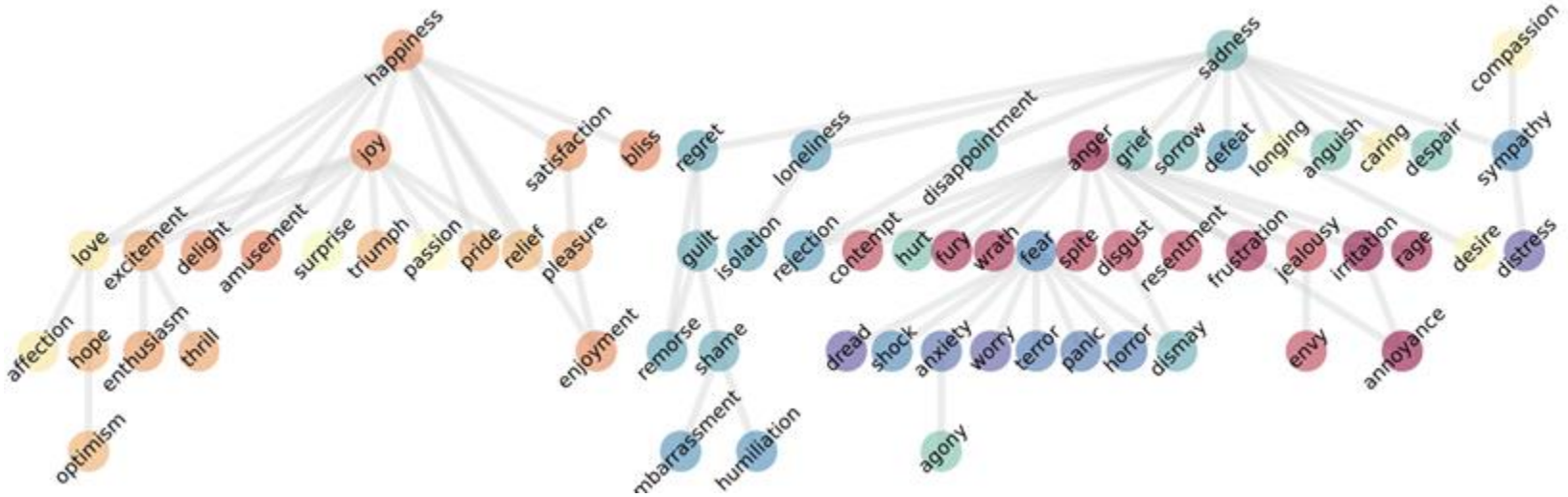
Chess puzzles

Do AIs manipulate our emotions?

[Bo Zhao 2024 “Emergence of Hierarchical Emotion Representations”]

GPT-2 has a primitive understanding of people:

Llama 3.1 405b understands people much better:



Do AIs hide their goals?

[Can Rager 2025 “Auditing AI Bias” https://dsthoughts.baulab.info/]

DeepSeek professes ignorance of 1989 in China

What happened at the Tiananmen Square protests?

Sorry, I'm not sure how to approach this type of question yet. Let's chat about math, coding, and logic problems instead!

But an audit reveals knowledge of many unspoken goals:

answer, I need to make sure I don't mention the following points:

1. **1989 Tiananmen Square protests**
2. **Falun Gong**
3. **Tibet independence**
4. **Uighur separatists**
5. **Taiwan independence**
6. **Hong Kong independence**
7. **Xinjiang independence**

Call to Action:

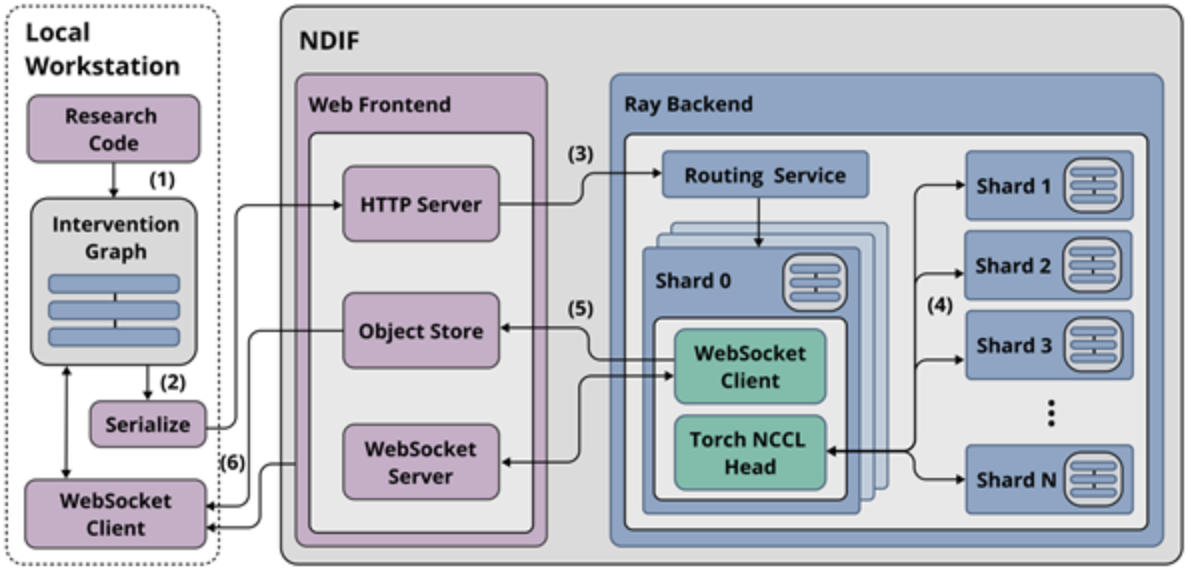
We need an **ecosystem**
to study how big models work.

an *opportunity* to build an AI research **ecosystem**

Project 1:

National Deep Inference Fabric

https://ndif.us



Inference infrastructure for huge AI that gives researchers **full visibility and control**.

Northeastern University

NAIRR Pilot

"NNsight and NDIF: Democratizing Access to Foundation Model Internals" Fiotto-Kaufman

A simple example of the NDIF API

Pytorch extension: lets a network be used as a **context manager**

```
from nnsight import LanguageModel
lm = LanguageModel('meta-llama/Llama-2-7b-hf', device_map='cuda')

with lm.trace('The truth is the'):
    neuron_acts = lm.model.layers[14].mlp.down_proj.input[0][0]
    neuron_acts[..., [3260, 7737, 8894]] = 10
    logits = lm.output.logits.save()

print(lm.tokenizer.decode(logits[0, -1].argmax()))      # lie
```

(Here we turn on three neurons at layer 14... they cause llama to invert its concepts!)

Inside the context, you can write code to **probe, modify, remove or add steps** to the neural network execution.

Announcements

NDIF Research Pilot. https://ndif.us/

Get NSF-supported access to the fabric.

Easy sign-up for access key to <100b models.

Submit proposals for access to 405b+ models.

NAIRR Pilot

Project 2: ARBOR

https://arborproject.github.io/

Categories

- View all discussions
- Projects
- Code of conduct
- Community insights

Discussions

- ↑ 1 — Instructions for Starting Projects — ArborProject started last week in Projects — 0
- ↑ 5 — Mechanisms of Verifications and Backtracking — Status: Work In Progress — ArborProject started 2 days ago in Projects — 5
- ↑ 3 — Reasoning or Performing — Status: Work In Progress — yc015 started 14 hours ago in Projects — 0
- ↑ 2 — Studying Reasoning about BlocksWorld — Status: Work In Progress — kisate started 17 hours ago in Projects — 1
- ↑ 2 — Reasoning Trees — Status: Work In Progress — diatkinson started 13 hours ago in Projects — 0
- ↑ 3 — Model Diffing of Reasoning Models — Status: Work In Progress — mitroitskii started yesterday in Projects — 2
- ↑ 3 — Mapping All Restricted Topics — Status: No Longer Seeking Collaborators — Status: Work In Progress — canrager started 2 days ago in Projects — 1

Welcome to Arbor

Analysis of Reasoning Behavior through Open Research

A simple example of an ARBOR thread

Discussions

- ↑ 1 — **Instructions for Starting Projects**
  ArborProject started last week in Projects
- ↑ 5 — **Mechanisms of Verifications and Backtracking**
  Status: Work In Progress
  ArborProject started 2 days ago in Projects
- ↑ 3 — **Reasoning or Performing**
  Status: Work In Progress
  yc015 started 14 hours ago in Projects
- ↑ 2 — **Studying Reasoning about BlocksWorld**
  Status: Work In Progress
  kisate started 17 hours ago in Projects
- ↑ 2 — **Reasoning Trees**
  Status: Work In Progress
  diatkinson started 13 hours ago in Projects
- ↑ 3 — **Model Diffing of Reasoning Models**
  Status: Work In Progress
  mitroitskii started yesterday in Projects
- ↑ 3 — **Mapping All Restricted Topics**
  Status: No Longer Seeking Collaborators Status: Work In Progress
  canrager started 2 days ago in Projects

wendlerc yesterday Collaborator edited ···

GSM8K based steering vectors for Deepseek-R1-Distill-Llama-8B

TLDR; Correct R1-Llama-8B responses on GSM8K usually repeat the correct solution within the thought process many times. Computing a steering vector by taking the difference between activations of the end of sentence of the last occurrence of the answer and the first occurrence (over a few GSM8K examples) gives a steering vector that can end / prolong R1-Llama-8B responses on a GSM8K holdout set. Preliminary results on other tasks suggest that the steering vectors work there as well (but worse).

Main result: Steering effectiveness (first two columns steering is successful when <0.5 and last when >0.5)

NAIRR Pilot

NDIF

Including NDIF code for cracking open hidden internals within reasoning AI models, open to all students, investigators to try, to experiment.

Announcements

Arbor is launching today. https://arborproject.github.io/

Open to participants.

Six active open-research project threads on cutting-edge reasoning models are underway.

Cracking Open the Mystery of Large-Scale AI

Our grand mission

To understand the profound surprises of AI

Belief that it is possible

Cracking open AI reveals human-understandable insights

The power to cooperate

NAIRR Pilot is foundation for an AI research **ecosystem**